1. Which of the formula below computes the **Accuracy** of a binary classifier formula?
    - (TP + TN) / (TP + FP + TN + FN)

2. In data preprocessing, which of the following are the objectives of the **aggregation** of attributes:
    - Obtain a less detailed scale
    - Reduce the number of attributes or distinct values
    - Reduce the variability of data

3. Which of the statements below best describe the strategy of **Apriori** in finding the frequent itemsets?
    - Evaluation of the support of the itemsets in an order such that uninteresting parts of the search space are pruned as soon as possible

4. Which of the following statements regarding the discovery of **association rules** is true?
    - The support of an itemset is anti-monotonic with respect to the composition of the itemset
    - The confidence of a rule can be computed starting from the supports of itemsets

5. Which of the following is a base hypothesis for a **Bayesian classifier**?
    - The attributes must be statistically independent inside each class

6. Which of the following **clustering method** is *NOT* based on distances between object?
    - Expectation Maximization

7. Which is the **confidence** of the rule A, C → B?
    - 50%

8. Which is the **confidence** of the rule B → E?
    - 33%

9. What is **cross validation**?
    - A technique to obtain a good estimation of the performance of a classifier when it will be used with data different from the training set

10. Which is the effect of the curse of **dimensionality**?
    - When the number of dimensions increases the Euclidean distance becomes less effective to discriminate between points in the space

11. After fitting **DBSCAN** with the default parameter values the results are: 0 clusters, 100% of noise points. Which will be your next trial?
    - Reduce the minimum number of objects in the neighborhood
    - Increase the radius of the neighborhood

12. Which of the following characteristics of data can reduce the effectiveness of **DBSCAN**?
    - Presence of clusters with different densities

13. Which of the following is a strength of the clustering algorithm **DBSCAN**?
    - Ability to separate outliers from regular data
    - Ability to find cluster with concavities

14. Which of the following is not a strengths point of **DBSCAN** with respect to K-Means?
    - The efficiency even in large datasets

15. Which of the statements below is true?
    - Sometimes DBSCAN stops to a configuration which does not include any cluster
    - Increasing the radius of the neighborhood can decrease the number of noise points
    - DBSCAN can give good performance when clusters have concavities

16. In a **decision tree**, an attribute which is used only in nodes near the leaves . . .
    - . . . gives little insight with respect to the target

17. In a **decision tree**, the number of object in a node . . .
    - . . . is smaller than the number of objects in its ancestor

18. A **Decision tree** is . . .
    - . . . A tree-structured plan of tests on single attributes to forecast the target

19. Which is **different** from the others? (Misclassification Error - Gini Index - Entropy)
    - Silhouette index

20. Which is **different** from the others? (DBSCAN – K-Means – Expectation Maximisation)
    - Decision Tree

21. Which is **different** from the others? (SVM – Decision Tree – Neural Network)
    - DBSCAN

22. Which is **different** from the others? (Apriori – K-Means – Expectation Maximisation)
    - Decision Tree

23. Which is the purpose of **discretization**?
    - Reduce the number of distinct values in an attribute, in order to put in evidence possible patterns and regularities

24. For each type of data choose the best suited **distance function**
    - High dimensional space → Manhattan distance
    - Vector space with real values → Euclidean distance
    - Boolean data → Jaccard coefficient
    - Vectors of terms representing documents → Cosine distance

25. What measure is maximized by the **Expectation Masimisation** algorithm for clustering?
    - The likelihood of a class label, given the values of the attributes of the example

26. In a dataset with D attributes, how many subsets of attributes should be considered for **feature selection** according to an exhaustive search?
    - O(2^D)

27. Which of the following in NOT an objective of **feature selection**?
    - Select the feature with higher range, which have more influence on the computations

28. Match the rule **evaluation formulas** with their names
    - $sup(A \cup C) - sup(A)sup(C)$ → Leverage
    - $\frac{conf(A \Rightarrow C)}{\sup(C)}$ → Lift
    - $\frac{sup(A \Rightarrow C)}{\sup(A)}$ → Confidence
    - $\frac{1-\sup(C)}{1-conf(A \Rightarrow C)}$ → Conviction

29. What is **Gini Index**?
    - An impurity measure of a dataset alternative to the Information Gain and to the Misclassification Index

30. Which of the following measure can be used as an alternative to the **Information Gain**?
    - Gini Index

31. The **Information Gain** is used to
    - Select the attribute which maximizes, for a given training set, the ability to predict the class value

32. In which mining activity the **Information Gain** can be useful?

- Classification

33. Given the two binary vectors below, which is their similarity according to the **Jaccard Coefficient**? (1000101101 – 1011101010)
    - 0.375
34. Which of the statements below is true?
    - Sometimes **K-Means** stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters

35. Which of the following characteristic of data can reduce the effectiveness of **K-Means**?
    - Presence of outliers

36. What does **K-Means** try to minimize?
    - The distortion, that is the sum of the squared distances of each point with respect to its centroid

37. In order to reduce the dimensionality of a dataset, which is the advantage of Multi Dimensional Scaling (**MDS**), with respect to Principal Component Analysis (**PCA**)?
    - MDS can be used with categorical data, provided that the matrix of the distance is available, while PCA is limited to vector spaces.

38. Which of the following is not a property of a **metric distance function**?
    - Boundedness

39. Which of the following preprocessing activities is useful to build a **Naïve Bayes** classifier if the independence hypothesis is violated?
    - Feature Selection

40. Which of the following statements is true?
    - The noise can generate outliers
    - Outliers can be due to noise

41. In data preparation which is the effect of the **normalization**?
    - Map all the numeric attributes to the same range, without altering the distribution, in order to avoid that attributes with large ranges have more influence

42. When developing a classifier, which of the following is a symptom of **overfitting**?
    - The error rate in the test set is much greater than the error rate in the training set

43. Which of the formulas below computes the **precision** of a binary classifier?
    - TP / (TP + FP)

44. Why do we **prune** a decision tree?
    - To eliminate parts of the tree where the decisions could be influenced by random effects

45. How does **pruning** work when generating frequent itemsets?
    - If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated

46. Which of the formulas below computes the **recall** of a binary classifier?
    - TP / (TP + FN)

47. Given the two binary vectors below, which is their similarity according to the **Simple Matching Coefficient**? (1000101101 – 1011101010)
    - 0.5

48. What is the **single linkage**?
    - A method to compute the distance between two sets of items, it can be used in hierarchical clustering

49. Which is the main purpose of the **smoothing** in Bayesian classification?
    - Classifying an object containing attribute values which are missing from some classes in the training set

50. Which is the main reason for the **standardization** of numeric attributes?
    - Map all the numeric attributes to a new range such that the mean is zero and the variance is one

51. What is the meaning of the statement: "the **support is antimonotone**"?
    - The support of an itemset never exceeds the support if its subsets

52. Which is the **support** of the rule A, C → B?
    - 20%

53. With reference to the total sum of squared errors and separation of a clustering scheme, which of the statements below is true?

They are strictly correlated, if, changing the clustering scheme, one increases, then the other decreases

54. What is the coefficient of determination R ?

Provide an index of goodness for a linear regression model

55. Which of the activities below is part of "Business Understanding" in the CRISP methodology?

Which are the resources available (manpower, hardware, software, ...)

56. Which is the main reason for the MinMax scaling (also known as "rescaling") of attributes?

Map all the numeric attributes to the same range, in order to prevent the attributes with higher range from having prevalent influence

57. What are the hyperparameters of a Neural Network? (Possibly non exhaustive

Hidden layers structure, Learning rate, Activation function, Number of epochs

58. How can we measure the quality of a trained regression model?

With a formula elaborating the difference between the forecast values and the true ones